PLOS Computational Biology

RESEARCH ARTICLE

Data-driven identification of biological systems using multi-scale analysis

Ismaila Muhammed[1☯], **Dimitris M. Manias**[1☯], **Dimitris A. Goussis**[2], **Haralampos Hatzikirou**[1,3,4*]

1 Department of Mathematics, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates, **2** Department of Mechanical Engineering, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates, **3** Biotechnology Center, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates, **4** Center for Information Services and High Performance Computing, Technische Universitat Dresden, Dresden, Germany

☯ These authors contributed equally as co-first authors.
* haralampos.hatzikirou@ku.ac.ae

Citation: Muhammed I, Manias DM, Goussis DA, Hatzikirou H (2025) Data-driven identification of biological systems using multi-scale analysis. PLoS Comput Biol 21(11): e1013193. https://doi.org/10.1371/journal.pcbi.1013193

Editor: William R. Cannon, Pacific Northwest National Laboratory, UNITED STATES OF AMERICA

Received: June 4, 2025

Accepted: October 24, 2025

Published: November 6, 2025

Data availability statement: All synthetic data and code necessary to reproduce the results of this study are publicly available in the github repository: https://github.com/drmitss/dd-multiscale. Here we do not use real biological data but only synthetic/simulated ones.

Abstract

Biological systems inherently exhibit multi-scale dynamics, making accurate system identification particularly challenging due to the complexity of capturing a wide time scale spectrum. Traditional methods capable of addressing this issue rely on explicit equations, limiting their applicability in cases where only observational data are available. To overcome this limitation, we propose a data-driven framework that integrates the Sparse Identification of Nonlinear Dynamics (SINDy) method, the multi scale analysis algorithm Computational Singular Perturbation (CSP) and neural networks (NNs). This framework allows the partition of the available dataset in subsets characterized by similar dynamics, so that system identification can proceed within these subsets without facing a wide time scale spectrum. Accordingly, when the full dataset does not allow SINDy to identify the proper model, CSP is employed for the generation of subsets of similar dynamics, which are then fed into SINDy. CSP requires the availability of the gradient of the vector field, which is estimated by the NNs. The framework is tested on the Michaelis-Menten model, for which various reduced models in analytic form exist at different parts of the phase space. It is demonstrated that the CSP-based data subsets allow SINDy to identify the proper reduced model in cases where the full dataset does not. In addition, it is demonstrated that the framework succeeds even in the cases where the available data set originates from stochastic versions of the Michaelis-Menten model. This framework is algorithmic, so system identification is not hindered by the dimensions of the dataset.

PLOS Computational Biology

Funding: This work was supported by the Volkswagen Stiftung for the "Life?" initiative (96732 to HH), Khalifa University of Science and Technology (RIG-2023-051 to HH), and the UAE-NIH Collaborative Research grant (AJF-NIH-25-KU to HH and DMM). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Biological systems often evolve across multiple time scales, posing major challenges for constructing accurate models directly from data. Traditional model reduction techniques require explicit equations and thus cannot be applied when only observational data are available. To address this, we developed a data-driven framework that combines Sparse Identification of Nonlinear Dynamics (SINDy), Computational Singular Perturbation (CSP) and neural networks (NNs). Our approach automatically partitions a dataset into subsets characterized by similar dynamics, allowing valid reduced models to be identified in each region. When SINDy fails to recover a global model from the full dataset, CSP -leveraging Jacobian estimates from NNs- successfully isolates dynamical regimes where SINDy can be applied locally. We validated this framework using the Michaelis-Menten biochemical model, which is known to admit multiple reduced models in different regions of the phase space. Our method consistently identified the appropriate reduced dynamics, even when the data originated from stochastic simulations. Because our approach is algorithmic and equation-free, it is scalable to high-dimensional systems and robust to noise, offering a promising solution for data-driven model discovery in complex biological systems.

Introduction

Mathematical modeling in biological systems often relies on first-principles approaches, typically formulated as differential equations, to describe, predict, and analyze biological processes. Experimental data is also used to calibrate and refine these models [1]. However, the complex, nonlinear, and multi-scale nature of such systems presents significant challenges for deriving accurate governing equations solely through traditional methods [2]. A key difficulty lies in the inability to capture the full spectrum of time scales that characterize the evolution of the system. Model reduction is a major approach to address this limitation, since it combines low dimensionality with the preservation of key dynamical features. Moreover, such reduction enables efficient analysis and interpretation [3]. Various methods have been developed for this purpose, mainly based on available governing equations. Such methods, like Quasi-Equilibrium (QE) [4], Quasi-steady-state Approximation (QSSA) [5], Computational Singular Perturbation (CSP) [6] and the method of invariant manifolds [7–9], decompose the set of variables into fast and slow components, by identifying low-dimensional manifolds and models that govern the long-term dynamics of the slow components.

Mathematical models can also be obtained by extracting dynamics directly from observational data [2]. A variety of data-driven methods have been developed to identify governing equations. These methods employ techniques such as sparsity promotion, symbolic regression, or machine learning to address the multi-scale nonlinear dynamics. Sparse Identification of Nonlinear Dynamics (SINDy) [10] is a prominent method that identifies sparse models by selecting a minimal set of nonlinear functions to capture system dynamics. Weak SINDy [11] improves robustness

against noisy and sparse data, especially beneficial for multi-scale systems with variable noise levels. More recently, iNeural SINDy has enhanced this framework by integrating neural networks and using an integral formulation to better handle noisy and sparse datasets [12]. Symbolic regression methods like PySR [13] use evolutionary algorithms to discover closed-form equations, making them suitable for capturing nonlinear behaviors, even in the presence of sparse data. Additionally, Physics-Informed Neural Networks (PINNs) [14] incorporate physical laws into their structure, enabling accurate model predictions from limited data. ARGOS, another symbolic regression method, uses evolutionary algorithms to discover interpretable, sparse models, building on methods like SINDy and PySR while introducing improvements for handling complex systems [15]. Dynamic Mode Decomposition (DMD) [16] and its extended version, EDMD [17], are effective for identifying principal modes and predicting system evolution from relatively sparse data. However, DMD does not recover explicit equations and its accuracy can be limited when the data do not sufficiently capture the underlying dynamics, particularly in multi-scale systems [18].

When the construction of governing equations is impractical or unfeasible, Jacobian estimation methods provide an efficient alternative for analyzing local system behavior. These methods focus on characterizing the local stability or linearization near equilibrium points, which is particularly useful for multi-scale and nonlinear systems. Techniques such as automatic differentiation through NNs [19] approximate the Jacobian matrix from data, allowing the assessment of system sensitivity without requiring a full model. The Lyapunov method, typically used to evaluate equilibrium stability, can also be applied to approximate the Jacobian matrix [20,21]. Additionally, kernel-based approaches like Gaussian processes estimate Jacobians by fitting smooth functions to data, providing an efficient means to analyze local dynamics in highly nonlinear settings [22]. The Koopman operator theory offers another strategy by linearizing nonlinear systems, enabling finite-dimensional approximations for control and Jacobian estimation [23]. However, data sparsity remains a critical limitation, as sparse sampling, particularly in multi-scale systems, can lead to inaccurate derivative estimates and unreliable Jacobian matrices [18].

Fig 1 provides an overview of the most commonly used data-driven methods for full system identification and Jacobian matrix estimation, categorized by their requirements in terms of the number of variables and data points. Each method has inherent trade-offs, especially when dealing with high-dimensional systems or limited data. As the number of variables increases, the risk of overfitting also rises, particularly in the presence of sparse or noisy data. This challenge is amplified

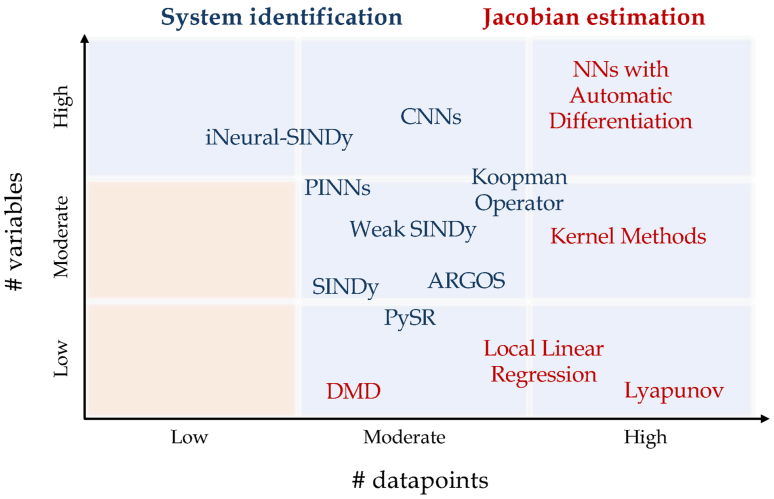

Fig 1. Existing data-driven methods. Schematic representation of existing data-driven methods for full system identification and Jacobian matrix estimation as a function of the number of variables and data points. Each method occupies a distinct region based on its data and dimensionality requirements. No methods can provide accurate estimations in the low variable-low datapoints region.

https://doi.org/10.1371/journal.pcbi.1013193.g001

in multi-scale systems, where critical dynamical features span multiple scales, making it difficult to accurately capture system behavior. Neural network-based methods are particularly data-intensive and prone to performance degradation when data is sparse, while other methods that decompose data into modes often require a dense set of observations to capture the full range of dynamics [18,24]. These challenges underscore the need for hybrid or regularization techniques that can effectively handle multi-scale systems, balancing model complexity with data availability.

A novel hybrid framework is introduced here that employs time scale decomposition for model identification in biological systems. By integrating the weak formulation of SINDy [10,11], CSP [25] and NN-based Jacobian estimation [19], the approach identifies algorithmically regions within a given dataset, in which valid reduced models can be constructed. In addition, the approach allows for extracting algorithmically mechanistic insight from multi-scale biological datasets. It is demonstrated that the proposed approach succeeds in cases where existing methodologies fail when applied to the full dataset. It should be emphasized that the focus of the present work is the identification of governing dynamics in the form of differential equations. The framework is therefore not intended for individual-based stochastic descriptions, which require different identification strategies.

The merits of the proposed framework are demonstrated on the basis of the Michaelis-Menten (MM) model, a widely used and well-established model in biological research for studying enzyme-substrate interactions and reaction kinetics [26,27]. Due to its simplicity and interpretability, the MM model is extensively used in Systems Biology, biochemistry, and computational biology [28,29]. Despite its straightforward formulation that allows analytical tractability, the model's nonlinear interactions and multi-scale dynamics present challenges for conventional system identification, making it a suitable benchmark for evaluating data-driven reduction methods [30,31]. Moreover, the MM model is of particular interest for studying and demonstrating the proposed framework, as in specific regions of its parametric space the system exhibits a shift in its slow dynamics, which causes existing methods to fail in correctly identifying the dynamics and providing valid reduced models. To further demonstrate the applicability of the framework to systems of higher dimension and more complex dynamical behavior, additional examples are presented in S1 Appendix and S2 Appendix: one system exhibiting a single transition from fast to slow dynamics, and another exhibiting two transitions, from fast to slow and from slow to slower.

Materials and methods

Sparse Identification of Nonlinear Dynamics (SINDy)

SINDy, introduced by Brunton et al. [10], is a framework used to identify dynamical systems from time series data using sparse regression. The main idea of SINDy is the assumption that the dynamics of the system can be represented as a sparse combination of candidate functions, making it computationally efficient and interpretable. Weak SINDy, a variant of SINDy was proposed by Schaeffer H. [32], to address the challenges of noise and and irregularly sampled datasets. This method reformulates the system identification problem in a weak form by integrating against a set of test functions, thereby reducing sensitivity to noise and enabling the use of coarsely sampled data. Despite its robustness to noise, the applicability of Weak SINDy is limited to models that conform to predefined functional forms, typically involving linear or weakly nonlinear relationships among variables. It struggles to identify complex models characterized by strong nonlinear interactions or variable-dependent nonlinearity, thereby limiting its effectiveness in systems with intricate dynamical structures [32].

Data-driven Jacobian estimation

To perform time scale decomposition and model reduction using CSP, the Jacobian matrix of the system is a key component. When explicit dynamical equations are unavailable, we use a combination of Neural Ordinary Differential Equations (NODE) [33] and NNs [34] to estimate the Jacobian matrix from data. NODE, introduced by Chen et al. [33], provides a way of learning the vector field that governs the continuous-time evolution of a system from data using NNs with the

adjoint sensitivity method, which efficiently optimizes the parameters of the system through gradient-based techniques. Frederic et al. introduced a novel approach for training NNs to estimate the Jacobian matrix of an unknown multivariate function using only input-output data pairs $(\mathbf{x}, \mathbf{F}(\mathbf{x}))$ [34]. The method utilizes a loss function based on the nearest neighbor search and linear approximations within the sample data.

The Computational Singular Perturbation Method (CSP)

The CSP algorithm allows for the analysis of multi-scale systems of ordinary differential equations (ODEs), by enabling the local decomposition of the tangent space into fast and slow subspaces [35,36]. At leading order, the two subspaces can be approximated by the right eigenvectors of the Jacobian, allowing the resolution of the vector field onto fast and slow components [37,38]. The fast component vanishes over a short transient, implying the equilibration of fast processes and the emergence of constraints that define the *Slow Invariant Manifold* (SIM). On this manifold, the slow dynamics evolve under the influence of the slow component of the vector field [39,40]. The CSP-based algorithmic vector field decomposition allows system-level identification of dominant fast and slow processes and their influence on the system's behavior, independent of dimensionality or nonlinearity. CSP has been widely applied to multi-scale systems, such as chemical reaction networks [41–45] and combustion configurations [46–48], but also in oscilating biological systems [49,50], population dynamics [51] and in systems describing cancer evolution [52], where time-scale separation presents significant analytical and computational challenges. A mathematical description of the methodology is provided in S3 Appendix.

Proposed framework

Fig 2 presents a schematic overview of the proposed framework. Given a time series dataset, SINDy -or any other such method- is first applied, when feasible, to directly identify the governing equations. In cases where SINDy fails, either due to noise, data sparsity or low data volume, NODE [33] is utilized to provide a uniform and dense vector field that is subsiquently used in a NN [34] to estimate the Jacobian matrix. Then, the estimated Jacobian matrix is considered for CSP to be applied without the need for explicit governing equations. The eigenvectors and eigenvalues of the estimated Jacobian serve as leading-order approximations of the system's corresponding fast/slow directions and time scales. Using the CSP theory and diagnostic tools [49,53,54], the dataset is then analyzed to identify regions where valid models can be constructed. These regions enable the partitioning of the dataset into subsets, each corresponding to a different dynamical regime. Finally, SINDy is applied to each subset independently to derive region-specific models that accurately capture the dynamics within each regime.

Results

We demonstrate the application of our methodology to the multi-scale Michaelis-Menten model, which can be simplified to three different reduced models that are valid in different domains of the phase and parameter spaces; the standard Quasi-Steady-State Approximation (sQSSA), the reverse Quasi-Steady-State Approximation (rQSSA) and the Partial Equilibrium Approximation (PEA) [31,50]. Although the model was introduced more than a century ago [26], its popularity keeps increasing [55] due to its significance in many biological and medical contexts [56].

The reduced Michaelis-Menten models

Equation-based model reduction. The MM reaction mechanism describes the interaction between a substrate S and an enzyme E, forming a reversible complex C. The complex C subsequently undergoes an irreversible reaction, yielding a product P and releasing the enzyme E, which can then participate in another cycle of substrate binding [26]. This process

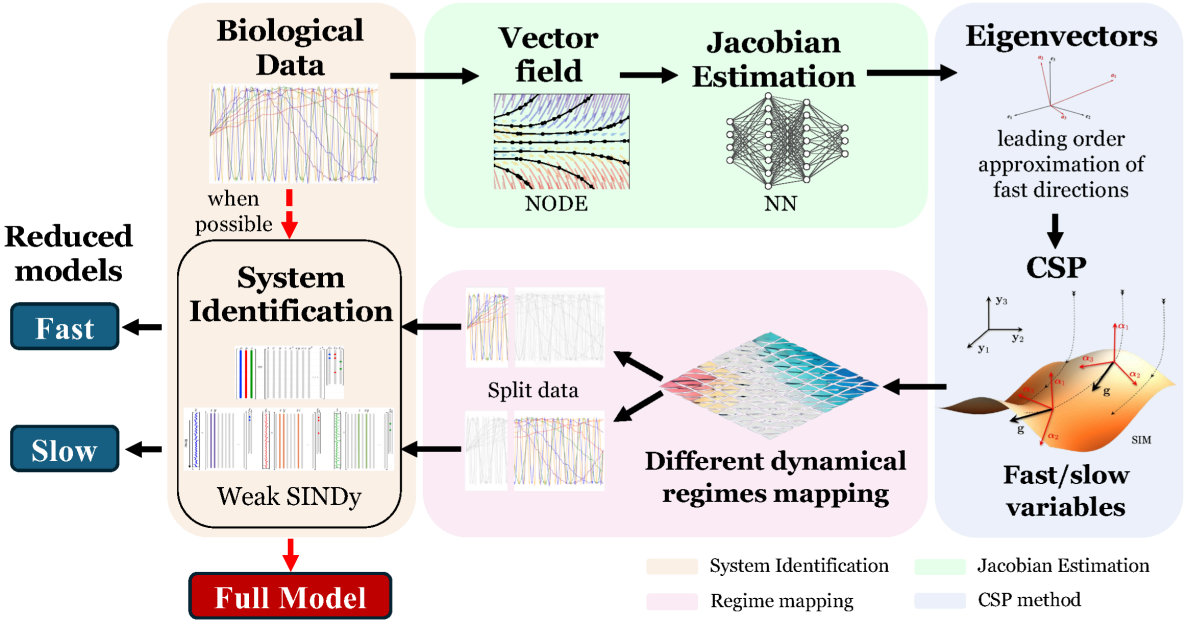

Fig 2. Framework. The methodology proposed to identify regions for valid models when system identification methodologies fail. The methodology starts with the availability of a biological dataset, which is directed either to a direct system identification methodology, when it is possible, or to the framework proposed here, when the first is not possible.

https://doi.org/10.1371/journal.pcbi.1013193.g002

can be represented as follows:

$$S + E \underset{k_b}{\overset{k_f}{\rightleftharpoons}} C \xrightarrow{k_2} E + P \tag{1}$$

where k_f (L mol$^{-1}s^{-1}$) and k_b (s^{-1}) denote the forward and reverse rate constants of the enzyme-substrate complex formation, respectively, while k_2 (s^{-1}) represents the catalytic rate constant, also referred to as the turnover number. According to the law of mass action, the MM mechanism is modeled by the following set of ordinary differential equations (ODEs):

$$\begin{aligned} \frac{d[C]}{dt} &= k_f[E][S] - k_b[C] - k_2[C] & \frac{d[S]}{dt} &= -k_f[E][S] + k_b[C] \\ \frac{d[E]}{dt} &= -k_f[E][S] + k_b[C] + k_2[C] & \frac{d[P]}{dt} &= k_2[C] \end{aligned} \tag{2}$$

Here, the square brackets denote the concentrations of the respective chemical species. Assuming the system is closed, where initially no product or complex is present, i.e., $[C](0) = [P](0) = 0$, the conservation relations $[E] + [C] = [E](0)$ and $[S] + [C] + [P] = [S](0)$ hold ($[E](0)$ and $[S](O)$ are the initial enzyme and substrate concentrations), the system in Eq (2) simplifies to:

$$\frac{d}{dt}\begin{bmatrix} c \\ s \end{bmatrix} = \mathbf{S}_1 R^1 + \mathbf{S}_2 R^2 = \begin{bmatrix} +1 \\ -1 \end{bmatrix} (k_{1f}(e_0 - c)\, s - k_{1b} c) + \begin{bmatrix} -1 \\ 0 \end{bmatrix} k_2 c \tag{3}$$

where the simplified symbols $c = [C]$, $s = [S]$ and $e_0 = [E](0)$ have been used. Assuming R^{1f} and R^{1b} are the rates of the forward and backward directions of the enzyme-substrate complex formation reaction, we set the bidirectional reaction

rate $R^1 = R^{1f} - R^{1b} = k_{1f}(e_0 - c)s - k_{1b}c$ and $R^2 = k_2 c$ is the rate of the catalytic reaction. $\mathbf{S}_1$ and $\mathbf{S}_2$ are the stoichiometric vectors of the two reactions. It should be noted that the origin, $(c, s) = (0, 0)$, is the only equilibrium point of the system.

The multi-scale nature of the two-dimensional system in Eq (3) is manifested through the significant difference in magnitude between the two time scales, τ_1 and τ_2 ($\tau_1 < \tau_2$), which govern the system's dynamics. These time scales can be approximated by the inverse modulo of the eigenvalues λ_+ and λ_- of the Jacobian matrix J of the two-dimensional ODE system described in Eq (3) [35,42,57]:

$$\lambda_\pm = \frac{-k_2}{2}\left(\frac{1+\mu+\nu}{\mu}\right)\left(1 \pm \sqrt{1-4\epsilon}\right) \tag{4}$$

where

$$\mu = \frac{K}{K_R + s}, \quad \nu = \frac{e_0 - c}{K_R + s}, \quad \epsilon = \frac{\mu\nu}{(1+\mu+\nu)^2} \tag{5}$$

and $K_R = k_{1b}/k_{1f}$ is the dissociation constant, $K = k_2/k_{1f}$ is the Van Slyke-Culen constant and $K_M = K_R + K$ is the Michaelis-Menten constant [31]. The value of ϵ is indicative to the gap that develops between the two eigenvalues (and corresponding time scales) and defines the stiffness of the system; i.e., $\tau_1/\tau_2 = f(\epsilon)$, so that $\tau_1/\tau_2 \ll 1$ when $\epsilon \ll 1$ and $\tau_1/\tau_2 \to 1$ when $\epsilon \to 1/4$ [58,59]. Both μ and ν are non-dimensional and non-negative and quantify the relative availability and initial distribution of enzyme and substrate in dimensionless form, determining the model's dynamic behavior and the separation between fast and slow reaction processes. The analytical expression of the eigenvalues in terms of μ and ν allows us identify regions in the μ-ν plane where different reduced models can be constructed [60].

Fig 3 displays the regions of validity of different reduced models. In particular, the region of a valid sQSSA model is highlighted with pink, the region of a valid rQSSA model is highlighted with green and the region of valid Partial Equilibrium Approximation (PEA) model is highlighted with shaded blue lines. The regions are separated by the lines $\nu = 1 + \mu$ and $\mu = 1 + \nu$. Along the $\mu < 1$ narrow neighborhood of the $\nu = 1 + \mu$ line (solid), indicated by fading to white color, no valid QSSA model can be obtained and only the PEA model is valid. Along the $\mu > 1$ part of the $\nu = 1 + \mu$ line (solid) $\epsilon \to 1/4$, so

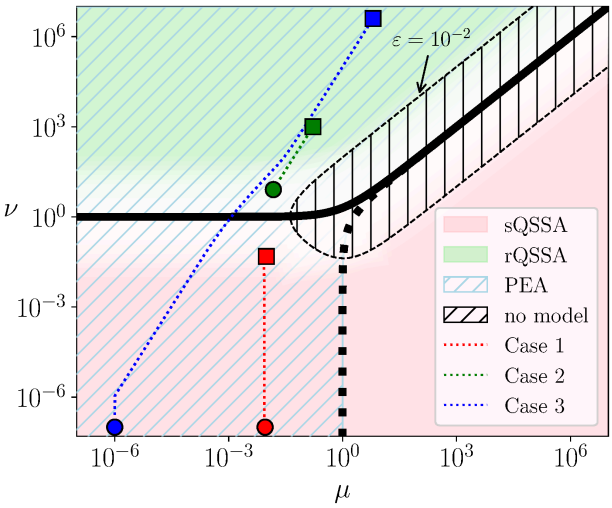

Fig 3. Regions of valid reduced models in the μ-ν plane. The regions of validity of sQSSA (pink: $\nu \ll \mu + 1$), rQSSA (green : $\nu \gg \mu + 1$) and PEA (shaded blue: $\mu \ll \nu + 1$) in the μ-ν plane, along with the trajectories that are analyzed (see Table 2 for the related parameters and ICs). Circles and squares denote the starting and ending point of each trajectory, respectively. The thick solid and dotted lines denote $\nu = 1 + \mu$ and $\mu = 1 + \nu$, respectively. The thin dashed curve denotes the points at which $\epsilon = 10^{-2}$ and encapsulates the shaded region where $\epsilon > 10^{-2}$. Reduced models in this region are of low accuracy [60].

https://doi.org/10.1371/journal.pcbi.1013193.g003

that no reduced model can be constructed there since $\tau_1/\tau_2 \to 1$. Moving away from this part, ϵ progressively decreases, as it is indicated by the $\epsilon = 10^{-2}$ dashed curve [31,60]. Table 1 displays the reduced models that are valid in the specified portions of the $\mu - \nu$ plane; i.e., the sQSSA model (c is considered fast), the rQSSA model (s is considered fast) and the PEA model (the bi-directional reaction $S + E \leftrightarrow C$ is in equilibrium, $R^{1f} \approx R^{1b}$). It is shown in Table 1 that the PEA model simplifies to the sQSSA model when $\nu \ll 1$ and to the rQSSA model when $\nu \gg 1$.

The fact that the full model in Eq (3) is valid throughout the $\mu - \nu$ plane and the three reduced models in Table 1 are valid in portions of this plane, will be the basis for the assessment of the proposed framework for system identification. In particular, datasets originating from the three trajectories shown in Fig 3 will form the starting point of the identification process. As shown in the figure, one trajectory is located in the region where the sQSSA and the PEA models are valid (Case 1), another is located in the region where the rQSSA and the PEA models are valid (Case 2) and a third trajectory is located in a region that includes the domains of validity of the sQSSA/PEA models (first part of the trajectory), the PEA model (middle part) and the rQSSA/PEA models (last part) (Case 3). The parameters and initial conditions for the three trajectories are displayed in Table 2.

The validity of the reduced models in Table 1 is demonstrated in Fig 4, where profiles of the three trajectories considered here obtained with the full and appropriate reduced model are compared. It is shown that an excellent agreement is obtained.

The simplification of the PEA model to the two QSSA models in Case 3 is demonstrated in Fig 5 that compares the solution of the PEA model with the solutions provided by the two QSSA models of Table 1. The sQSSA-based profiles of c and s in the left panel and the rQSSA profiles in the right panel are denoted by dashed lines, while the PEA profiles in both panels are denoted by crosses. It is evident that the sQSSA solution closely follows the original trajectory in the first region (pink, as in Fig 3), where the sQSSA is valid, but begins to diverge as the system transitions into the second region (green, as in Fig 3), where only rQSSA is valid. Conversely, the rQSSA solution initially deviates from the original trajectory in the first region, where rQSSA is not valid, but progressively aligns with it upon entering the second region, where rQSSA is valid.

Table 1. The algebraic relation that defines the SIM and the differential equation(s) that governs the flow on the SIM, according to sQSSA, rQSSA and PEA.

Region	Algebraic Eq (SIM)	Differential Eq (reduced system)
sQSSA	$(e_0 - c)\,s - K_R c - Kc = 0$ $\left[c = \frac{e_0 s}{K_M + s}\right]$	$\frac{ds}{dt} = -k_2 c$ $\left[\frac{dc}{dt} = -\frac{k_2 (e_0 - c)^2 c}{K_M e_0}\right]$
rQSSA	$(e_0 - c)\,s - K_R c = 0$ $\left[s = \frac{K_R c}{e_0 - c}\right]$	$\frac{dc}{dt} = -k_2 c$ $\left[\frac{ds}{dt} = -\frac{k_2 (K_R + s) s}{K_R}\right]$
PEA	$(e_0 - c)\,s - K_R c - \frac{1}{1+\nu} Kc = 0$	$\frac{dc}{dt} = -\frac{\nu}{1+\nu} k_2 c$ $\frac{ds}{dt} = -\frac{1}{1+\nu} k_2 c$

In brackets are the traditional forms of the relations that approximate the SIM and the differential equation that governs the evolution of the fast variable in the sQSSA and rQSSA context.

https://doi.org/10.1371/journal.pcbi.1013193.t001

Table 2. Parameters and initial conditions for simulation of the cases considered.

Case	k_{1f}	k_{1b}	k_2	e_0	$s(0)$	$c(0)$
Case 1	10	100	1	0.5	1	0.045045
Case 2	3	3	0.5	1000	10.11111	910
Case 3	1000	0.01	0.1	10	10	9.999999

The choice of the values is based on the analysis in [61] and [60]. Initial conditions are obtained on the *SIM* for each case.

https://doi.org/10.1371/journal.pcbi.1013193.t002

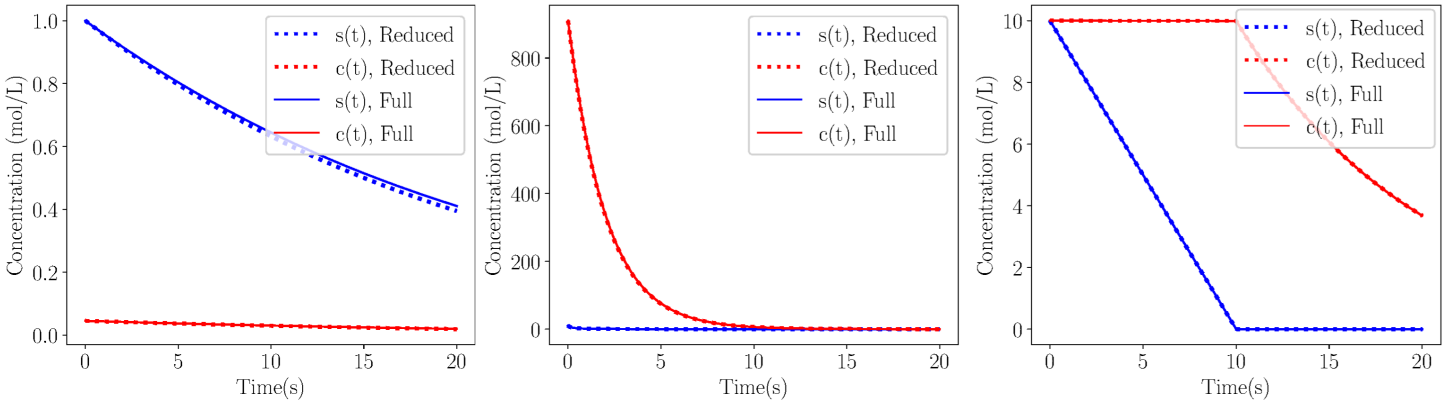

Fig 4. **System Solutions.** Solution comparison between full (solid) and reduced (dotted) Michaelis-Menten simulations on the SIM, for Case 1 - sQSSA (left), Case 2 - rQSSA (middle) and Case 3 - PEA (right). Parameters and initial conditions on Table 2.

https://doi.org/10.1371/journal.pcbi.1013193.g004

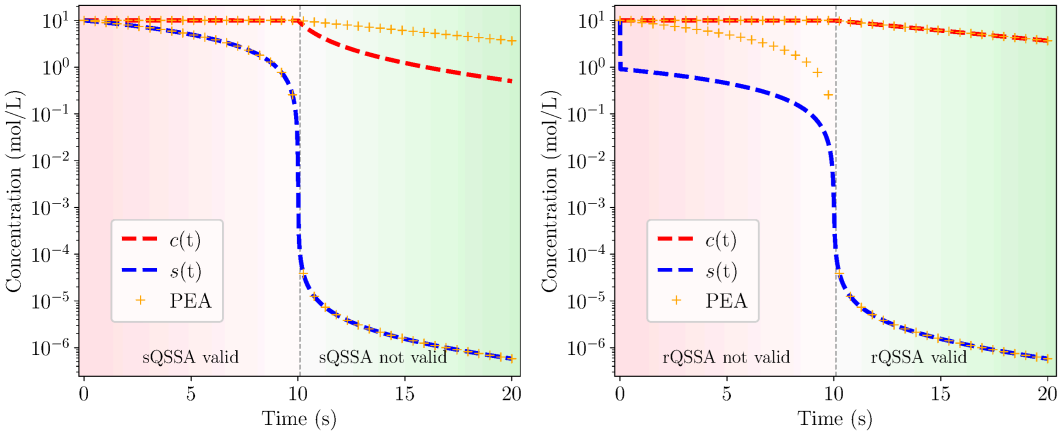

Fig 5. **Validity comparison.** Validity comparison between the sQSSA (left) and rQSSA (right) and PEA solutions of Case 3 (see Table 2 for parameters and ICs).

https://doi.org/10.1371/journal.pcbi.1013193.g005

Data-driven reduced model identification using SINDy. In order to assess the validity of the proposed method for system identification, datasets were produced from the profiles shown in Fig 4 from both the full and reduced models, by considering a relatively sparse and uniform grid of n=100 datapoints; a limit close to where SINDy might fail.

The solutions of each case, both of the full and the reduced models, were used as datasets in Weak SINDy for system identification. The `WeakPDELibrary` from `PySINDy` was employed to construct the weak-form feature library, which avoids explicit derivative computation by integrating the governing equations against selected test functions. The library was built from polynomial functions up to cubic order, including univariate terms (x, x^2, x^3) and mixed products (xy), with corresponding custom names for clarity. The spatiotemporal grid was set to the simulation time vector, and the weak formulation used $K = 2000$ integration points to ensure stable integral estimates. For the sparse regression step, we used the SR3 optimizer with an ℓ_1 thresholding rule, and a maximum of 1000 iterations. We systematically varied the sparsity threshold over the range [0.01,10] to examine model sensitivity.

The results are displayed in Table 3. The *coefficient of determination*, denoted as R^2, was used to evaluate how well the identified models fit the data. R^2 measures the proportion of the variance in the dependent variable (typically the

Table 3. Comparison of the right-hand side expressions between the ground truth and the identified models, using data obtained from the deterministic full (top) and reduced (bottom) models.

	Case	Ground Truth	Identified Model	R^2
Full Model	Case 1	$\dot{c} = 5s - 101c - 10sc$ $\dot{s} = -5s + 100c + 10sc$	$\dot{c} = 5.004s - 101.073c - 10.003sc$ $\dot{s} = -5.003s + 100.07c + 10.003sc$	0.9991 0.9990
	Case 2	$\dot{c} = 3000s - 3.5c - 3sc$ $\dot{s} = -3000s + 3c + 3sc$	$\dot{c} = 2999.915s - 3.5c - 3sc$ $\dot{s} = -3000.007s + 3c + 3sc$	1.0 1.0
	Case 3	$\dot{c} = 100000s - 0.11c - 10000sc$ $\dot{s} = -100000s + 0.01c + 10000sc$	$\dot{c} = -1000004.014s - 0.11c - 10000.401sc$ $\dot{s} = -100004.038s + 0.01c + 10000.404sc$	0.9959 0.9975
Reduced Model	Case 1	$\dot{c} = -0.050cc + 0.198c^2 - 0.198c^3$ $\dot{s} = -1c$	$\dot{c} = -0.050c + 0.198c^2 - 0.198c^3$ $\dot{s} = -0.999c$	0.9999 1.0
	Case 2	$\dot{c} = -0.5c$ $\dot{s} = -0.5s - 0.5s^2$	$\dot{c} = -0.5c$ $\dot{s} = -0.5s - 0.5s^2$	0.9996 1.0
	Case 3	$\dot{c} = -\frac{\nu}{1+\nu}0.1c$ $\dot{s} = -\frac{1}{1+\nu}0.1c$	$\dot{c} = -5944980.763s + 0.620c + 594498.076sc$ $\dot{s} = 5945000.197s - 0.720c - 594500sc$	-2.5614 -4.3303

The coefficient of determination (R^2) indicates the accuracy of the Weak SINDy reconstruction.

https://doi.org/10.1371/journal.pcbi.1013193.t003

system's dynamics) that is predictable from the independent variables (the terms in the identified governing equations). The R^2 metric was chosen as it is the standard measure used in the SINDy framework to assess the accuracy of identified models, thereby ensuring consistency and comparability with existing studies.

The coefficient of determination is computed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{6}$$

where y_i indicates the actual observed data (here, the derivative of the state, $[\dot{c}, \dot{s}]^\top$) and $\hat{y}_i$ indicates the predicted values from the SINDy model. The mean obsearved data is given by $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$. When $R^2 = 1$, the predicted model explains all the variance in the data.

The results in Table 3 indicate that Weak SINDy successfully identified the different models corresponding to the data generated from the full models. In particular, it accurately recovered the parameters of each system, as reflected by the high coefficient of determination values (R^2), which are consistently close to unity.

When applied to data from the reduced models, Weak SINDy accurately identified the reduced dynamics in the two QSSA cases. However, in Case 3 the method failed to recover the reduced model from the data. Instead, it returned a polynomial approximation resembling the structure of the full model, with completely different coefficients. This discrepancy is reflected in the large negative values of R^2, indicating that the identified model structure does not capture the underlying dynamics accurately.

This failure is attributed to the limitations of SINDy in constructing models with complex nonlinear terms that do not conform to its predefined candidate library. The PEA reduced model includes nonlinear terms of the form $1/(1+\nu)$ and $\nu/(1+\nu)$, where $\nu = (e_0 - c)/(K_R + s)$ is itself a nonlinear function of the state variables. Since SINDy relies on a fixed library of basis functions, typically comprising polynomials, trigonometric functions, or other simple expressions, it struggles to represent such composite nonlinearities. However, when SINDy is directly provided with functions that explicitly include the nonlinear terms of the numerators and denominators, it successfully captures the dynamics and correctly identifies the PEA reduced model [10]. Despite this, it remains impractical to know a priori the specific nonlinear terms required for accurate model identification in real-world applications, where the underlying functional forms are typically

unknown. This highlights a fundamental challenge when using SINDy for complex multi-scale systems with non-standard interactions.

Implementing our proposed method for Case 3

The framework proposed here addresses the challenges encountered by Weak SINDy in identifying a valid reduced model in Case 3 as follows. First, the NODE network [33] is employed to generate a dense and uniform vector field, needed for the accurate implementation of the NN-based Jacobian approximation. NODE is utilized in an unsupervised manner, as the objective is to infer the underlying dynamical system from the observed trajectories without explicit labels. The NODE model was implemented as a multilayer feedforward neural network with hidden layers of size [64,128,128,64] and ReLU activation functions. The network was trained to approximate the system dynamics by minimizing the mean squared error between the predicted and reference trajectories. Training was performed using the Adam optimizer with a learning rate of 10^{-3}, weight decay of 10^{-4}, and gradient clipping (maximum norm of 1.0) to ensure numerical stability. A step scheduler halved the learning rate every 500 epochs. The solver tolerances for trajectory integration were set to $\text{rtol} = 10^{-6}$ and $\text{atol} = 10^{-6}$, ensuring high accuracy in trajectory reconstruction. Unless otherwise stated, training was carried out for 2000 epochs without early stopping, although the framework includes a configurable patience-based stopping criterion. To evaluate the learned dynamics beyond the training interval, the NODE was further used to extrapolate trajectories at higher temporal resolution (up to 2000 points), from which dense vector fields and their derivatives were computed.

Subsequently, the NN methodology proposed in [34] is used to estimate the structure of the Jacobian matrix directly from the reconstructed vector field values, without requiring additional supervision. Hyperparameters, such as nearest neighbors, neighborhood radius and NN architecture, as well as training settings including optimizer and training parameters, are carefully tuned to ensure accurate and stable Jacobian estimation. In detail, the model consists of four hidden layers with 600, 600, 300, and 150 neurons, respectively, and employs the Swish activation function. Training was performed with a batch size of 64 for 150 epochs, using the Adam optimizer with a learning rate of 1×10^{-4}. To improve numerical stability, gradient constraints were applied through maximum-norm regularization. Prior to training, both the state variables and their derivatives were normalized to zero mean and unit variance, with an inverse transformation applied afterward to recover the physical scaling of the Jacobian. The loss function was designed to minimize the discrepancy between the true and NN-predicted differential increments across local neighborhoods of the dataset, thus ensuring that the estimated Jacobian matrices preserved the underlying vector field structure.

The eigenvectors of the Jacobian matrix are then employed by CSP to approximate the fast and slow directions in phase space. Additionally, the inverse of the eigenvalue moduli are utilized to approximate the characteristic time scales of the system. When providing a full dataset produced from the PEA model for Case 3, CSP analysis revealed that the dataset contains two distinct regions where vastly different dynamics prevail. In agreement to the results in Fig 3, CSP concluded that: (i) in the first region, the enzyme-substrate complex c exhibits fast dynamics, justifying the applicability of sQSSA; and (ii) in the second region, the substrate s transitions to the fast variable, making rQSSA a valid approximation. Furthermore, CSP explicitly identified the narrow region, where the transition between these two dynamical regimes is realized, allowing for an appropriate partitioning of the dataset in two parts, so that the data in each set are characterized by similar dynamics. Details of the CSP analysis are presented in S4 Appendix.

Following this dataset partitioning, each subset corresponding to a different QSSA model is independently processed using the Weak SINDy algorithm. The results, summarized in Table 4, demonstrate that our framework successfully identified the appropriate QSSA model within each region. During the transition period (white region in Fig 3 and Fig 5), where no valid QSSA model can be constructed but only a PEA model, Weak SINDy identified the PEA model only when the nonlinear terms in the model were added in the library. A comparison between the PEA and identified models in Table 4 might suggest that the zero derivatives of the fast variables are incorrect. However, a closer inspection reveals that these

Table 4. Comparison of the right-hand side expressions between the PEA and the identified models by the Weak SINDy, using the CSP-based split of the dataset of Case 3.

	Region	Full model	PEA model	Identified model	R^2
Case 3	Split 1 (sQSSA)	$\dot{c} = 10^5 s - 0.11c - 10^4 sc$	$\dot{c} = -\frac{\nu}{1+\nu} 0.1c$	$\dot{c} = 0$ $\dot{s} = -0.1c$	–1.1969 0.9999
	Split 2 (rQSSA)	$\dot{s} = -10^5 s + 0.01c + 10^4 sc$	$\dot{s} = -\frac{1}{1+\nu} 0.1c$	$\dot{c} = -0.1c$ $\dot{s} = 0$	1.0 –7.7547

Data obtained from the full model. R^2 indicates coefficient of determination of Weak SINDy.

https://doi.org/10.1371/journal.pcbi.1013193.t004

results are structurally consistent with the expected reduced dynamics. This is shown in Fig 6, which illustrates the temporal evolution of the terms $1/(1+\nu)$ and $\nu/(1+\nu)$ in the PEA model over the full data set $\nu \approx 0$, leading to . In the first period, which refers to the data subset where sQSSA is valid, Fig 6 shows that $1/(1+\nu) \approx 1$ and $\nu/(1+\nu) \approx 0$. In these limits, the PEA model simplifies to the identified model. In the second period, which refers to the data subset where rQSSA is valid, $1/(1+\nu) \approx 0$ and $\nu/(1+\nu) \approx 1$. Again, in these limits the PEA model simplifies to the identified model.

The negative values of R^2 in Table 4 arise when the derivatives of the fast variables are close to zero. In such cases, the variance of the corresponding time series is negligible, and the variance-based definition of R^2 produces misleadingly negative values. These values should therefore not be interpreted as incorrect dynamics, but rather as artifacts of the metric when applied to low-variance variables. Conversely, the negative R^2 values observed in Table 3 (PEA case) represent genuine model mismatch: in these cases, Weak SINDy produces a polynomial approximation that fails to capture the nonlinear structure of the PEA reduced dynamics. Thus, negative R^2 values can arise for two different reasons: (i) as artifacts in low-variance contexts, which do not imply model failure, and (ii) as indicators of the inability of Weak SINDy to represent certain nonlinear dynamics. This distinction clarifies how R^2 should be interpreted in multiscale system identification.

It was demonstrated here that, while Weak SINDy initially failed to identify a valid global model for the entire dataset, the proposed framework successfully partitioned the data into dynamically distinct regions. Within these regions, our

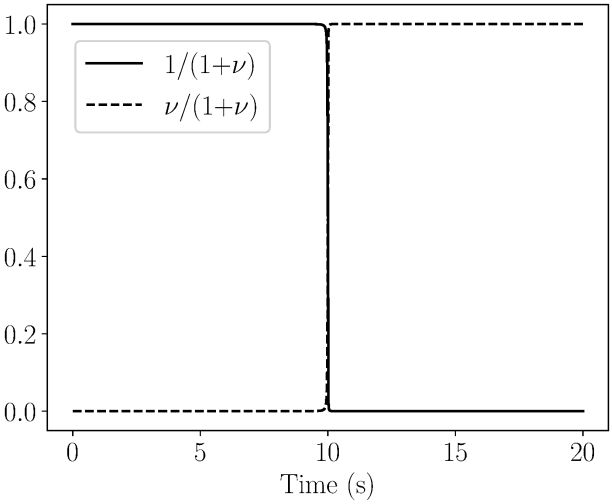

Fig 6. Evolution of $1/(1+\nu)$ and $\nu/(1+\nu)$. Evolution of $1/(1+\nu)$ (solid) and $\nu/(1+\nu)$ (dashed) for Case 3. In the first part where sQSSA is valid, $1/(1+\nu) \approx 1$ and $\nu/(1+\nu) \approx 0$, while in the second part where rQSSA is valid, $1/(1+\nu) \approx 0$ and $\nu/(1+\nu) \approx 1$.

https://doi.org/10.1371/journal.pcbi.1013193.g006

framework accurately recovered locally valid reduced models, emphasizing that the primary objective in a purely data-driven setting is to identify models that capture local dynamics rather than a single globally valid formulation.

Model identification for noisy data

In real-world biological applications, data contain inherent noise and may exhibit underlying time scale separation. This further complicates the identification of governing dynamics and the construction of valid models using methods such as SINDy, and even its robust variant, Weak SINDy. To simulate noise in the dynamics, we modify the deterministic system in Eq (3) by introducing stochastic perturbations in two forms: additive and multiplicative noise.

It is important to note that our identification strategy focuses on recovering dynamics in the form of differential equations. For this reason, noise was introduced through stochastic differential equations (SDEs), which are appropriate for mesoscopic and macroscopic regimes where mean-field descriptions remain valid. Alternative approaches, such as stochastic simulation algorithms that track individual reaction events (e.g., the Gillespie algorithm), are better suited for systems with very small molecule numbers. However, such cases require fundamentally different identification strategies and fall outside the scope of the present work.

Additive noise is introduced by adding a distrurbance in the right-hand side of the system with Gaussian noise independent of the vector field:

$$\frac{d}{dt}\begin{bmatrix} c \\ s \end{bmatrix} = \begin{bmatrix} +1 \\ -1 \end{bmatrix} (k_{1f}(e_0 - c)\, s - k_{1b} c) + \begin{bmatrix} -1 \\ 0 \end{bmatrix} k_2 c + \eta(t), \tag{7}$$

where $\eta(t) = [\eta_c(t), \eta_s(t)]^\top$ is a vector-valued Gaussian noise process with zero mean and standard deviation proportional to the signal magnitude:

$$\eta_i(t) \sim \mathcal{N}\left(0,\ \sigma^2\right), \qquad \text{with } \sigma = D \cdot \sqrt{\frac{1}{n}\sum x_i^2}, \quad i \in \{c, s\}, \tag{8}$$

x_i denotes the value of the respective variable at time t and D denotes the noise percent or strength.

Multiplicative noise is modeled as a stochastic perturbation that scales with the magnitude of the vector field:

$$\frac{d}{dt}\begin{bmatrix} c \\ s \end{bmatrix} = \left(\begin{bmatrix} +1 \\ -1 \end{bmatrix} (k_{1f}(e_0 - c)\, s - k_{1b} c) + \begin{bmatrix} -1 \\ 0 \end{bmatrix} k_2 c \right) \circ (\mathbf{1} + \xi(t)), \tag{9}$$

where $\circ$ denotes element-wise multiplication, $\mathbf{1}$ is a vector of ones matching the state dimension and $\xi(t) = [\xi_c(t), \xi_s(t)]^\top$ is a vector-valued Gaussian noise process with zero mean and standard deviation proportional to the signal magnitude:

$$\xi_i(t) \sim \mathcal{N}(0,\ \sigma^2), \qquad \text{with } \sigma = D \cdot \sqrt{\frac{1}{n}\sum x_i^2}, \quad i \in \{c, s\}. \tag{10}$$

In both cases, noise is introduced at the level of the vector field (i.e., the right-hand side of the ODEs), simulating realistic measurement or process noise often observed in biological data. Noisy datasets were generated for both the full and reduced models, derived from Eq (3) and Table 1, respectively.

The results obtained by applying Weak SINDy to the noisy datasets are summarized in Table 5. For Cases 1 and 2, Weak SINDy successfully identified reduced models, with the right-hand side of the fast variable equations correctly approximated as zero, consistent with the quasi-steady-state assumption. In contrast, the method failed to reconstruct any valid model for Case 3, which involves more complex multi-scale dynamics, under noisy conditions. Fig 7 illustrates the impact of additive (2%) and multiplicative (1%) noise on the system trajectories in Case 3. While additive noise did not significantly alter the trajectory, multiplicative noise led to substantial deviations from deterministic values, especially

Table 5. Performance of Weak SINDy on identifying the corresponding models of the data produced from the full and reduced stochastic models.

Case	2% Additive Noise		1% Multiplicative Noise	
	Full	Reduced	Full	Reduced
Case 1	✓*	✓*	✓*	✓*
Case 2	✓*	✓*	✓*	✓*
Case 3	✗	✗	✗	✗

✓* and ✓* refer to cases where the RHS of the fast variable's equation in the identified model is equal to zero.

https://doi.org/10.1371/journal.pcbi.1013193.t005

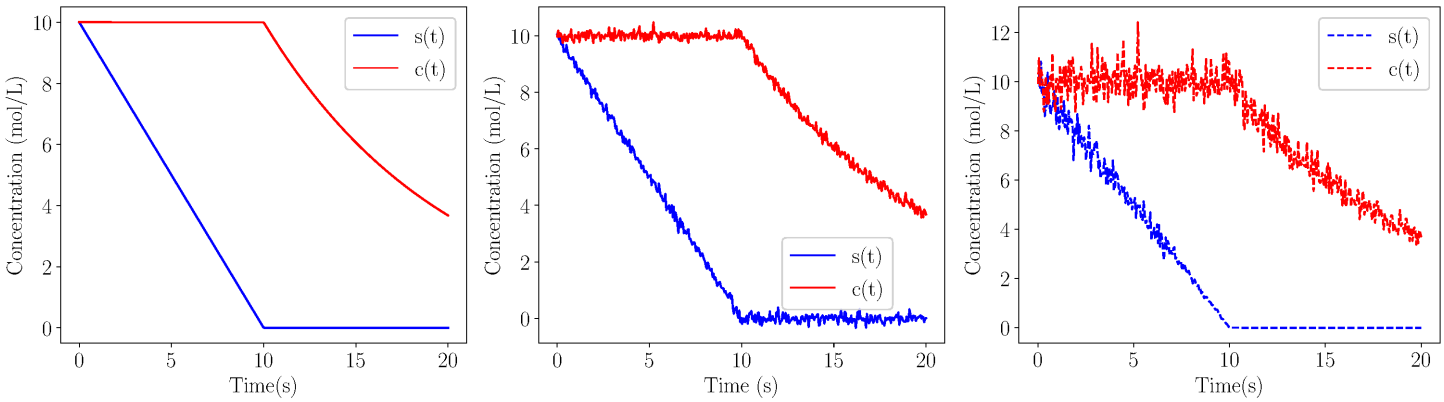

Fig 7. System evolution under noisy conditions. Temporal profiles of the variables reconstructed from NODE for the PEA case: deterministic data (left), data with 2% additive noise (middle) and data with 1% multiplicative noise (right).

https://doi.org/10.1371/journal.pcbi.1013193.g007

at higher magnitudes, preventing Weak SINDy from correctly identifying the underlying dynamics. This limitation underscores the challenge posed by the combined effects of multi-scale behavior and noise, emphasizing the necessity of employing the proposed framework for more robust model identification under realistic, noisy biological data scenarios.

Table 6 presents the results of applying Weak SINDy to both deterministic and noisy datasets of Case 3. In all scenarios, the method identified structurally valid reduced QSSA models, correctly setting the right-hand side of the fast variable to zero. Although the associated R^2 values are negative, this outcome stems from the vanishing variance of fast variable derivatives under QSSA assumptions, which limits the metric's reliability in such contexts. The consistent recovery of valid models across varying noise levels underscores the robustness of the proposed framework.

Table 6. Comparison of the right-hand side expressions between the PEA and the identified models, using the CSP-based split of the dataset of Case 3.

Region	Full model	PEA model	No Noise		Additive Noise		Mutl. Noise	
			Model	R^2	Model	R^2	Model	R^2
Split 1	$\dot{c} = 10^5 s - 0.11c - 10^4 sc$	$\dot{c} = -\frac{\nu}{1+\nu} 0.1c$	$\dot{c} = 0$	-1.1969	$\dot{c} = 0$	-1.1969	$\dot{c} = 0$	-1.1969
			$\dot{s} = -0.1c$	0.9999	$\dot{s} = -0.097c$	0.9973	$\dot{s} = -0.093c$	0.9827
Split 2	$\dot{s} = -10^5 s + 0.01c + 10^4 sc$	$\dot{s} = -\frac{1}{1+\nu} 0.1c$	$\dot{c} = -0.1c$	1.0	$\dot{c} = -0.099c$	0.9999	$\dot{c} = -0.099c$	0.9999
			$\dot{s} = 0$	-7.7547	$\dot{s} = 0$	-7.7547	$\dot{s} = 0$	-7.7547

Data obtained from the deterministic full model. R^2 indicates coefficient of determination of Weak SINDy.

https://doi.org/10.1371/journal.pcbi.1013193.t006

Fig 8 compares the deterministic solutions (solid lines) of the PEA model to those identified by Weak SINDy (dotted lines) after partitioning the dataset into the two subsets in which sQSSA (top row) and rQSSA (bottom row) are valid. The three columns correspond to no noise (left), additive noise (middle), and multiplicative noise (right). All data were smoothed using the NODE framework to enable accurate vector field reconstruction. Across all noise levels, the identified trajectories for substrate s(t) (blue) and complex c(t) (red) align closely with the PEA trajectories, demonstrating the effectiveness of the framework in identifying valid reduced models even in the presence of noise. The results of the CSP analysis of the stochastic cases are presented in S5 Appendix.

A summary of the results from all stochastic simulations is displayed in Table 7, where the performance of Weak SINDy in identifying the full and reduced models is assessed and compared with that of the deterministic simulations. The deterministic models include results for both full and reduced models, while the stochastic models consider cases with additive and multiplicative noise at different noise levels (2% for additive and 1% for multiplicative). The results indicate that for the deterministic cases, Weak SINDy successfully identifies both the full and reduced models in Cases 1 and 2, whereas for Case 3, it fails to identify a reduced model when data from the reduced model is used. However, after partitioning the dataset into distinct regions (Case 3 Split 1 and Case 3 Split 2), corresponding to valid sQSSA and rQSSA models, the proposed framework successfully identifies the correct reduced models in each sub-region.

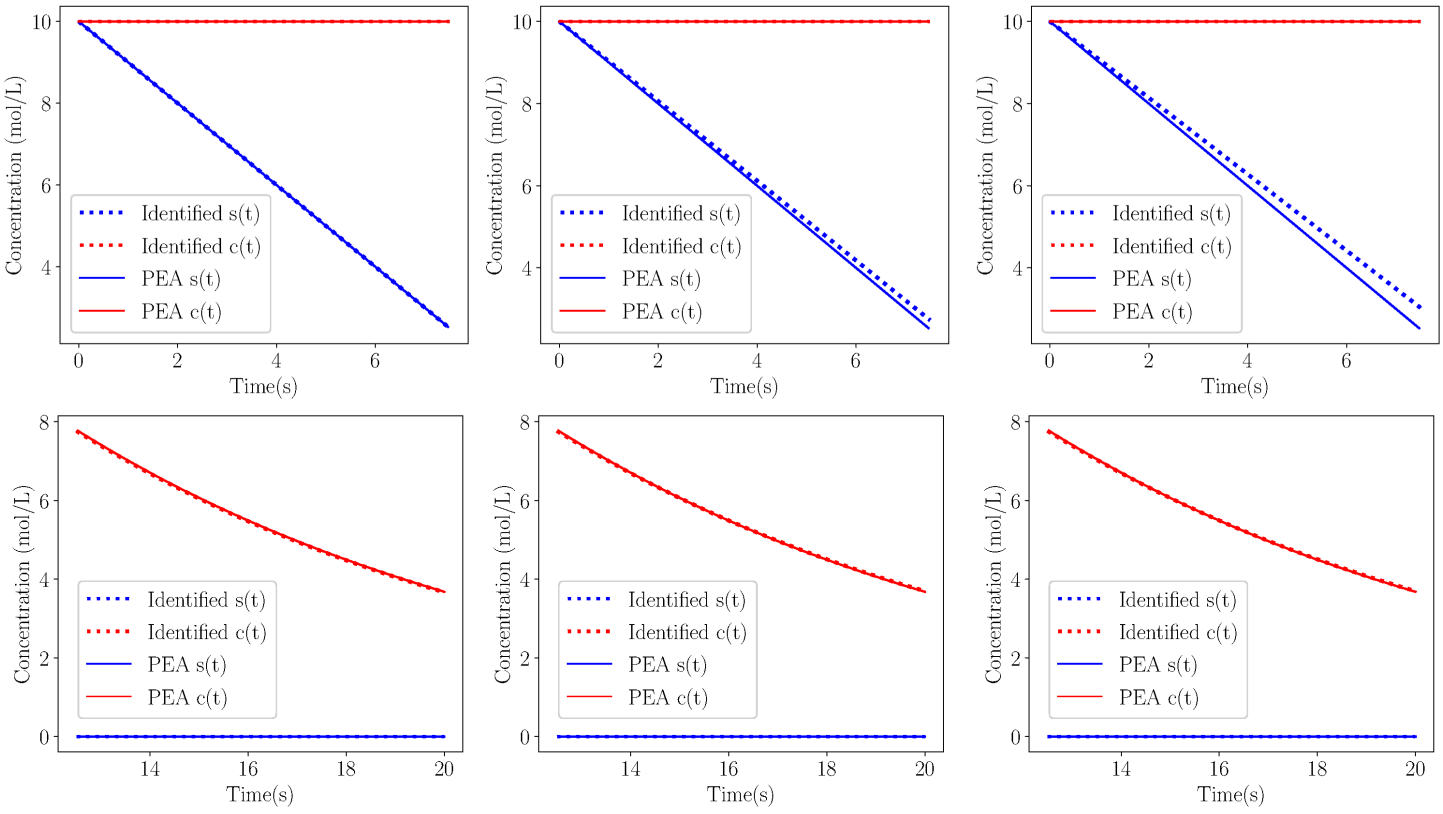

Fig 8. Comparisons. Solution comparison between the actual deterministic (solid) and identified from Weak SINDy (dotted) models of Michaelis-Menten simulations of Case 3 on the SIM, after spliting the data into regions where sQSSA (top) and rQSSA (bottom) models are valid. The data for the cases of no noise (left), additive (middle) and multiplicative noise (right) have been smoothened by the NODE. Parameters and initial conditions on Table 2.

https://doi.org/10.1371/journal.pcbi.1013193.g008

Table 7. Performance of Weak SINDy on identifying the corresponding models of the data produced from the full and reduced deterministic and stochastic models.

Case	Deterministic model		Stochastic model			
	Full	Reduced	2% Additive Noise		1% Multiplicative Noise	
			Full	Reduced	Full	Reduced
Case 1	✓	✓	✓*	✓*	✓*	✓*
Case 2	✓	✓	✓*	✓*	✓*	✓*
Case 3	✓	✗	✗	✗	✗	✗
Case 3 Split 1	✓*	-	✓*	-	✓*	-
Case 3 Split 2	✓*	-	✓*	-	✓*	-

✓* and ✓* refer to cases where the RHS of the fast variable's equation in the identified model is equal to zero.

https://doi.org/10.1371/journal.pcbi.1013193.t007

Discussion

We applied our proposed framework to the Michaelis-Menten system, a canonical model that captures the nonlinear and multi-scale nature of biological dynamics. Despite its simplicity, the model offers three analytically tractable time-scale regimes, namely sQSSA, rQSSA and PEA, serving as ideal test cases for validating data-driven methods. Through deterministic and stochastic simulations, we demonstrated that Weak SINDy can identify valid reduced models under appropriate conditions. However, it fails when applied to full datasets exhibiting transitions between regimes or containing non-standard nonlinearities. By integrating NODE-based vector field reconstruction, NN-based Jacobian estimation and CSP analysis, the framework successfully partitioned the data into dynamically consistent subregions, enabling accurate model identification across all cases, including noisy ones.

An important aspect of the proposed framework is its applicability to experimental datasets, where measurements are often noisy, sparse, or incomplete. By leveraging NODE for vector field reconstruction and the NN-driven Jacobian estimation, the framework can be applied even when only partial observations of the system are available. This enables the identification of latent dynamics and the recovery of reduced models under conditions that more closely resemble real-world biological data. The expected benefits include improved robustness to measurement noise and enhanced reliability in capturing multiscale dynamics and regime transitions that conventional system identification methods fail to resolve.

The summary of all simulations, shown in Table 7, confirms the strengths and limitations of Weak SINDy under different conditions. For deterministic data, the method recovered both full and reduced dynamics in Cases 1 and 2 but failed to reconstruct the PEA reduced model in Case 3. Upon partitioning the Case 3 dataset into separate sQSSA and rQSSA regimes, the proposed framework correctly identified the reduced dynamics. This success persisted under noisy conditions, provided that the data were preprocessed with NODE. In cases where the fast variable's derivative was near-zero, negative R^2 values appeared. As previously discussed, such values should not be interpreted as failures but rather as artifacts of the QSSA structure and the limitations of variance-based metrics.

The data in Case 3 were generated from simulations of the full Michaelis-Menten system, not from an already reduced model. The Michaelis-Menten model was deliberately chosen not only because its dynamics are analytically well understood -making it easier for the reader to follow the framework and its validation- but also because of its characteristic behavior in the PEA region, where the slow dynamics undergo a shift. Due to this property, SINDy and similar methods often fail to correctly identify the underlying dynamics, providing a meaningful test case for our framework. The main scope of this work is to address the inability of SINDy and related approaches to identify valid models in the presence of noise, a condition inherent to real biological data. For demonstration, we used the deterministic model to clearly illustrate the underlying dynamics, while applying the same full dataset under both deterministic and noisy conditions to ensure consistency. Furthermore, to demonstrate applicability beyond two-dimensional systems, we analyzed in Supplementary Sects S4 and S5 two three-dimensional models: one exhibiting a simple transition from fast to slow dynamics, and

another undergoing a double transition, from fast to slow and from slow to slower. This design highlights that the methodology is general and not restricted to simple systems, but rather provides a principled approach applicable to datasets of higher dimensionality and diverse dynamical behavior.

One limitation of the proposed method is its sensitivity to data quality. High noise levels or poor signal-to-noise ratios can obscure the underlying dynamics, particularly in systems governed by multiple time scales. Multiplicative noise, in particular, leads to significant distortion at high variable magnitudes, hampering the accuracy of the vector field and Jacobian estimation. While NODE can smooth noisy data to an extent, its effectiveness depends on the quality and density of the input data.

Another important limitation concerns the modeling framework itself. Our approach is explicitly designed to identify effective models in the form of differential equations, and therefore noise was incorporated at the level of SDEs. This choice is appropriate for mesoscopic and macroscopic regimes, where ODE and SDE descriptions are valid representations of the underlying dynamics. In contrast, in regimes with very small molecule numbers, individual-based stochastic simulations (e.g., Gillespie-type methods) provide a more realistic description, but such cases require fundamentally different identification strategies. Consequently, the present framework should be regarded as targeting the parameter ranges where differential equation models are appropriate, consistent with the overarching goals of this work.

The performance of the NN components, including NODE and the Jacobian estimator, also hinges on careful tuning of hyperparameters. The number of nearest neighbors, architecture, learning rates, and training epochs all influence the accuracy of the estimated Jacobian. Incorrect configurations may lead to poor identification of fast and slow directions, undermining the CSP analysis. While our framework performed robustly under controlled conditions, real-world applications will likely require additional strategies for hyperparameter optimization and model validation.

Data sparsity presents another significant challenge. Sparse or short trajectories may not provide sufficient information for reliable vector field reconstruction, even with NODE. In such cases, preprocessing steps such as interpolation may be necessary to increase the sampling density. However, over-interpolation may introduce artifacts, necessitating a balance between data augmentation and preservation of the original dynamics.

To overcome sparsity and enhance model identification in limited-data settings, we propose incorporating Dynamic Mode Decomposition (DMD). This method can generate synthetic data from the latent space, enriching the available trajectory without requiring additional measurements. By expanding the data coverage, DMD supports more accurate vector field reconstruction and Jacobian estimation, thereby improving the reliability of the subsequent CSP and SINDy analyses. Also the DMD can serve as a comparison for the identified model predictions, when the ground truth is unkwown.

Extending the proposed framework to high-dimensional systems represents a critical direction for future research. Although CSP is inherently dimension-agnostic and theoretically remains applicable, the scalability of neural estimators poses practical challenges, as both computational cost and data requirements increase with dimensionality. Adapting the framework to real-world, high-dimensional datasets -such as those arising in systems biology, neuroscience, or clinical monitoring- will demand advances in both computational efficiency and data utilization. Possible strategies include incorporating dimensionality-reduction techniques such as DMD to extract dominant modes of variability before neural estimation, as well as employing specialized neural architectures tailored to high-dimensional data, such as convolutional or graph-based networks. These approaches can mitigate computational complexity while preserving the essential multiscale structure of the dynamics. Given the complexity of biological systems, it is essential to develop integrative approaches that unify machine learning with mechanistic modeling [62–66]. Our method contributes to this goal, offering a foundation for scalable, data-driven multiscale analysis. Incorporating physics-informed constraints or domain-specific priors into the neural architecture offers a promising strategy to enhance generalization, reduce data demands, and broaden applicability across complex biological domains.

Supporting information

S1 Appendix. Application to a 3-dim stochastic model with one transition from fast to slow.
(PDF)

S2 Appendix. Application to a 3-dim model with two transitions from fast to slow and slow to slower.
(PDF)

S3 Appendix. CSP Methodology.
(PDF)

S4 Appendix. CSP diagnostics of the deterministic models.
(PDF)

S5 Appendix. CSP diagnostics of the stochastic model of Case 3.
(PDF)

Author contributions

Conceptualization: Haralampos Hatzikirou.

Data curation: Ismaila Muhammed, Dimitris M. Manias.

Formal analysis: Ismaila Muhammed, Dimitris M. Manias, Dimitris A. Goussis, Haralampos Hatzikirou.

Funding acquisition: Haralampos Hatzikirou.

Investigation: Ismaila Muhammed, Dimitris M. Manias, Dimitris A. Goussis, Haralampos Hatzikirou.

Methodology: Dimitris A. Goussis, Haralampos Hatzikirou.

Software: Ismaila Muhammed, Dimitris M. Manias.

Supervision: Haralampos Hatzikirou.

Validation: Dimitris A. Goussis.

Visualization: Ismaila Muhammed, Dimitris M. Manias.

Writing – original draft: Ismaila Muhammed, Dimitris M. Manias.

Writing – review & editing: Dimitris A. Goussis, Haralampos Hatzikirou.

References

1. Murari A, Peluso E, Lungaroni M, Gaudio P, Vega J, Gelfusa M. Data driven theory for knowledge discovery in the exact sciences with applications to thermonuclear fusion. Sci Rep. 2020;10(1):19858. https://doi.org/10.1038/s41598-020-76826-4 PMID: 33199734
2. Prokop B, Gelens L. From biological data to oscillator models using SINDy. iScience. 2024;27(4):109316. https://doi.org/10.1016/j.isci.2024.109316 PMID: 38523784
3. Southern J, Pitt-Francis J, Whiteley J, Stokeley D, Kobashi H, Nobes R, et al. Multi-scale computational modelling in biology and physiology. Prog Biophys Mol Biol. 2008;96(1–3):60–89. https://doi.org/10.1016/j.pbiomolbio.2007.07.019 PMID: 17888502
4. Gorban A. Model reduction in chemical dynamics: slow invariant manifolds, singular perturbations, thermodynamic estimates, and analysis of reaction graph. Current Opinion in Chemical Engineering. 2018;21:48–59. https://doi.org/10.1016/j.coche.2018.02.009
5. Semenov S, Starov V, Rubio RG, Velarde MG. Computer simulations of quasi-steady evaporation of sessile liquid droplets. In: Starov V, Prochazka K, editors. Trends in colloid and interface science XXIV. Berlin, Heidelberg: Springer; 2011. p. 115–20.
6. Lam SH, Goussis DA. Understanding complex chemical kinetics with computational singular perturbation. Symposium (International) on Combustion. 1989;22(1):931–41. https://doi.org/10.1016/s0082-0784(89)80102-x

7. Roussel MR, Fraser SJ. Geometry of the steady-state approximation: perturbation and accelerated convergence methods. The Journal of Chemical Physics. 1990;93(2):1072–81. https://doi.org/10.1063/1.459171
8. Singh S, Powers JM, Paolucci S. On slow manifolds of chemically reactive systems. The Journal of Chemical Physics. 2002;117(4):1482–96. https://doi.org/10.1063/1.1485959
9. Gorban AN, Karlin IV. Method of invariant manifold for chemical kinetics. Chemical Engineering Science. 2003;58(21):4751–68. https://doi.org/10.1016/j.ces.2002.12.001
10. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc Natl Acad Sci U S A. 2016;113(15):3932–7. https://doi.org/10.1073/pnas.1517384113 PMID: 27035946
11. Messenger DA, Bortz DM. Weak sindy for partial differential equations. J Comput Phys. 2021;443:110525. https://doi.org/10.1016/j.jcp.2021.110525 PMID: 34744183
12. Forootani A, Goyal P, Benner P. A robust SINDy approach by combining neural networks and an integral form. arXiv preprint 2023. https://arxiv.org/abs/230907193
13. Cranmer M. Interpretable machine learning for science with PySR and SymbolicRegression. arXiv preprint 2023. https://arxiv.org/abs/2305.01582
14. Patsatzis DG, Russo L, Siettos C. Slow invariant manifolds of fast-slow systems of ODEs with physics-informed neural networks. SIAM J Appl Dyn Syst. 2024;23(4):3077–122. https://doi.org/10.1137/24m1656402
15. Egan K, Li W, Carvalho R. Automatically discovering ordinary differential equations from data with sparse regression. Commun Phys. 2024;7(1):20.https://doi.org/10.1038/s42005-023-01516-2
16. SCHMID PJ. Dynamic mode decomposition of numerical and experimental data. J Fluid Mech. 2010;656:5–28. https://doi.org/10.1017/s0022112010001217
17. Williams MO, Kevrekidis IG, Rowley CW. A Data–driven approximation of the koopman operator: extending dynamic mode decomposition. J Nonlinear Sci. 2015;25(6):1307–46. https://doi.org/10.1007/s00332-015-9258-5
18. Wu Z, Brunton SL, Revzen S. Challenges in dynamic mode decomposition. J R Soc Interface. 2021;18(185):20210686. https://doi.org/10.1098/rsif.2021.0686 PMID: 34932929
19. Latremoliere F, Narayanappa S, Vojtechovsky P. Estimating the Jacobian matrix of an unknown multivariate function from sample values by means of a neural network. arXiv preprint 2022. https://arxiv.org/abs/2204.00523
20. Oku M, Aihara K. On the covariance matrix of the stationary distribution of a noisy dynamical system. NOLTA. 2018;9(2):166–84. https://doi.org/10.1587/nolta.9.166
21. Varando G, Hansen NR. Graphical continuous Lyapunov models. In: Conference on Uncertainty in Artificial Intelligence. PMLR; 2020. p. 989–98.
22. Kanagawa M, Hennig P, Sejdinovic D, Sriperumbudur BK. Gaussian processes and kernel methods: a review on connections and equivalences. 2018. https://arxiv.org/abs/1807.02582
23. Brunton SL, Brunton BW, Proctor JL, Kutz JN. Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control. PLoS One. 2016;11(2):e0150171. https://doi.org/10.1371/journal.pone.0150171 PMID: 26919740
24. D'Amario V, Srivastava S, Sasaki T, Boix X. The data efficiency of deep learning is degraded by unnecessary input dimensions. Front Comput Neurosci. 2022;16:760085. https://doi.org/10.3389/fncom.2022.760085 PMID: 35173595
25. Lam S, Goussis D, Konopka D. Time-resolved simplified chemical kinetics modelling using computational singular perturbation. In: 27th Aerospace Sciences Meeting; 1989. p. 575.
26. Michaelis L, Menten ML. Die kinetik der invertinwirkung. Biochem Z. 1913;49(2):333–69.
27. Segel IH. Enzyme kinetics: behavior and analysis of rapid equilibrium and steady state enzyme systems. vol. 115. Wiley: New York; 1975.
28. Shin S, Chae SJ, Lee S, Kim JK. Beyond homogeneity: assessing the validity of the Michaelis-Menten rate law in spatially heterogeneous environments. PLoS Comput Biol. 2024;20(6):e1012205. https://doi.org/10.1371/journal.pcbi.1012205 PMID: 38843305
29. Lim R, Martin TLP, Chae J, Kim WJ, Ghim C-M, Kim P-J. Generalized Michaelis-Menten rate law with time-varying molecular concentrations. PLoS Comput Biol. 2023;19(12):e1011711. https://doi.org/10.1371/journal.pcbi.1011711 PMID: 38079453
30. Gunawardena J. Some lessons about models from Michaelis and Menten. Mol Biol Cell. 2012;23(4):517–9. https://doi.org/10.1091/mbc.E11-07-0643 PMID: 22337858
31. Patsatzis DG, Goussis DA. Algorithmic criteria for the validity of quasi-steady state and partial equilibrium models: the Michaelis-Menten reaction mechanism. J Math Biol. 2023;87(2):27. https://doi.org/10.1007/s00285-023-01962-0 PMID: 37432484
32. Schaeffer H. Learning partial differential equations via data discovery and sparse optimization. Proc Math Phys Eng Sci. 2017;473(2197):20160446. https://doi.org/10.1098/rspa.2016.0446 PMID: 28265183
33. Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud DK. Neural ordinary differential equations. In: Advances in Neural Information Processing Systems. 2018. https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html
34. Latremoliere F, Narayanappa S, Vojtechovsky P. Estimating the Jacobian matrix of an unknown multivariate function from sample values by means of a neural network; 2022. http://arxiv.org/abs/2204.00523.
35. Lam SH, Goussis DA. Understanding complex chemical kinetics with computational singular perturbation. Symposium (International) on Combustion. 1989;22(1):931–41. https://doi.org/10.1016/s0082-0784(89)80102-x

36. Hadjinicolaou M, Goussis DA. Asymptotic solution of Stiff PDEs with the CSP method: the reaction diffusion equation. SIAM J Sci Comput. 1998;20(3):781–810. https://doi.org/10.1137/s1064827596303995
37. Zagaris A, Kaper HG, Kaper TJ. Analysis of the computational singular perturbation reduction method for chemical kinetics. J Nonlinear Sci. 2004;14(1):59–91. https://doi.org/10.1007/s00332-003-0582-9
38. Diamantis DJ, Mastorakos E, Goussis DA. H2/air autoignition: the nature and interaction of the developing explosive modes. Combustion Theory and Modelling. 2015;19(3):382–433.
39. Fenichel N. Geometric singular perturbation theory for ordinary differential equations. Journal of Differential Equations. 1979;31(1):53–98. https://doi.org/10.1016/0022-0396(79)90152-9
40. Kaper TJ. Systems theory for singular perturbation problems. In: Analyzing multiscale phenomena using singular perturbation methods: American Mathematical Society short course, Baltimore, Maryland, 1999. 85.
41. Goussis DA, Maas U. Model reduction for combustion chemistry. In: Echekki T, Mastorakos E, editors. Turbulent Combustion Modeling. London: Springer; 2011. p. 193–220.
42. Zagaris A, Kaper HG, Kaper TJ. Analysis of the computational singular perturbation reduction method for chemical kinetics. J Nonlinear Sci. 2004;14(1):59–91. https://doi.org/10.1007/s00332-003-0582-9
43. Valorani M, Creta F, Goussis DA, Lee JC, Najm HN. An automatic procedure for the simplification of chemical kinetic mechanisms based on CSP. Combustion and Flame. 2006;146(1–2):29–51. https://doi.org/10.1016/j.combustflame.2006.03.011
44. Khalil AT, Manias DM, Tingas E-Al, Kyritsis DC, Goussis DA. Algorithmic analysis of chemical dynamics of the autoignition of NH3–H2O2/air mixtures. Energies. 2019;12(23):4422. https://doi.org/10.3390/en12234422
45. Manias DM, Patsatzis DG, Kyritsis DC, Goussis DA. NH3 vs. CH4 autoignition: a comparison of chemical dynamics. Combustion Theory and Modelling. 2021;25(6):1110–31. https://doi.org/10.1080/13647830.2021.1890835
46. Tingas EA, Gkantonas S, Mastorakos E, Goussis D. The mechanism of propagation of NH3/air and NH3/H2/air laminar premixed flame fronts. International Journal of Hydrogen Energy. 2024;78:1004–15.
47. Rabbani S, Manias DM, Kyritsis DC, Goussis DA. Chemical dynamics of the autoignition of near-stoichiometric and rich methanol/air mixtures. Combustion Theory and Modelling. 2021;26(2):289–319. https://doi.org/10.1080/13647830.2021.2012260
48. Valorani M, Najm HN, Goussis DA. CSP analysis of a transient flame-vortex interaction. Combustion and Flame. 2003;134(1–2):35–53. https://doi.org/10.1016/s0010-2180(03)00067-1
49. Goussis DA, Najm HN. Model reduction and physical understanding of slowly oscillating processes: the circadian cycle. Multiscale Model Simul. 2006;5(4):1297–332. https://doi.org/10.1137/060649768
50. Patsatzis DG, Goussis DA. A new Michaelis-Menten equation valid everywhere multi-scale dynamics prevails. Math Biosci. 2019;315:108220. https://doi.org/10.1016/j.mbs.2019.108220 PMID: 31255632
51. Manias DM, Patsatzis DG, Goussis DA. Time scale dynamics of COVID-19 pandemic waves: the case of Greece. arXiv preprint. 2023. https://arxiv.org/abs/231207260
52. Patsatzis DG. Algorithmic asymptotic analysis: extending the arsenal of cancer immunology modeling. J Theor Biol. 2022;534:110975. https://doi.org/10.1016/j.jtbi.2021.110975 PMID: 34883121
53. Manias DM, Tingas EAI, Frouzakis CE, Boulouchos K, Goussis DA. The mechanism by which CH2O and H2O2 additives affect the autoignition of CH4/air mixtures. Combustion and Flame. 2016;164:111–25. https://doi.org/10.1016/j.combustflame.2015.11.008
54. Manias DM, Goldman RN, Goussis DA. Physical insights from complex multiscale non-linear system dynamics: identification of fast and slow variables. Communications in Nonlinear Science and Numerical Simulation. 2025;148:108858. https://doi.org/10.1016/j.cnsns.2025.108858
55. Cornish-Bowden A. One hundred years of Michaelis–Menten kinetics. Perspectives in Science. 2015;4:3–9. https://doi.org/10.1016/j.pisc.2014.12.002
56. Deichmann U, Schuster S, Mazat J-P, Cornish-Bowden A. Commemorating the 1913 Michaelis-Menten paper Die Kinetik der Invertinwirkung: three perspectives. FEBS J. 2014;281(2):435–63. https://doi.org/10.1111/febs.12598 PMID: 24180270
57. Maas U, Pope SB. Simplifying chemical kinetics: intrinsic low-dimensional manifolds in composition space. Combustion and Flame. 1992;88(3–4):239–64. https://doi.org/10.1016/0010-2180(92)90034-m
58. Zagaris A, Kaper HG, Kaper TJ. Fast and slow dynamics for the computational singular perturbation method. Multiscale Model Simul. 2004;2(4):613–38. https://doi.org/10.1137/040603577
59. Kaper HG, Kaper TJ, Zagaris A. Geometry of the computational singular perturbation method. Math Model Nat Phenom. 2015;10(3):16–30. https://doi.org/10.1051/mmnp/201510303
60. Patsatzis DG, Goussis DA. A new Michaelis-Menten equation valid everywhere multi-scale dynamics prevails. Math Biosci. 2019;315:108220. https://doi.org/10.1016/j.mbs.2019.108220 PMID: 31255632
61. Patsatzis DG, Goussis DA. Algorithmic criteria for the validity of quasi-steady state and partial equilibrium models: the Michaelis-Menten reaction mechanism. J Math Biol. 2023;87(2):27. https://doi.org/10.1007/s00285-023-01962-0 PMID: 37432484
62. Shojaee P, Weinholtz E, Schaadt NS, Feuerhake F, Hatzikirou H. Biopsy location and tumor-associated macrophages in predicting malignant glioma recurrence using an in-silico model. NPJ Syst Biol Appl. 2025;11(1):3. https://doi.org/10.1038/s41540-024-00478-7 PMID: 39779740

63. Hatzikirou H. Combining dynamic modeling with machine learning can be the key for the integration of mathematical and clinical oncology: comment on “Improving cancer treatments via dynamical biophysical models” by M. Kuznetsov, J. Clairambault, V. Volpert. Phys Life Rev. 2022;40:1–2. https://doi.org/10.1016/j.plrev.2022.01.002 PMID: 35085921

64. Metzcar J, Jutzeler CR, Macklin P, Köhn-Luque A, Brüningk SC. A review of mechanistic learning in mathematical oncology. Front Immunol. 2024;15:1363144. https://doi.org/10.3389/fimmu.2024.1363144 PMID: 38533513

65. Mascheroni P, López Alfonso JC, Kalli M, Stylianopoulos T, Meyer-Hermann M, Hatzikirou H. On the impact of chemo-mechanically induced phenotypic transitions in gliomas. Cancers (Basel). 2019;11(5):716. https://doi.org/10.3390/cancers11050716 PMID: 31137643

66. Benzekry S, Mastri M, Nicolò C, Ebos JML. Machine-learning and mechanistic modeling of metastatic breast cancer after neoadjuvant treatment. PLoS Comput Biol. 2024;20(5):e1012088. https://doi.org/10.1371/journal.pcbi.1012088 PMID: 38701089